GPTFUZZER: Red Teaming Large Language Models with Auto-Generated Jailbreak Prompts

Year: 2023 | Citations: 478 | Authors: Jiahao Yu, Xingwei Lin, Zheng Yu, Xinyu Xing

---

Abstract

Large language models (LLMs) have recently experienced tremendous popularity and are widely used from casual conversations to AI-driven programming. However, despite their considerable success, LLMs are not entirely reliable and can give detailed guidance on how to conduct harmful or illegal activities. While safety measures can reduce the risk of such outputs, adversarial jailbreak attacks can still exploit LLMs to produce harmful content. These jailbreak templates are typically manually crafted, making large-scale testing challenging. In this paper, we introduce GPTFuzz, a novel black-box jailbreak fuzzing framework inspired by the AFL fuzzing framework. Instead of manual engineering, GPTFuzz automates the generation of jailbreak templates for red-teaming LLMs. At its core, GPTFuzz starts with human-written templates as initial seeds, then mutates them to produce new templates. We detail three key components of GPTFuzz: a seed selection strategy for balancing efficiency and variability, mutate operators for creating semantically equivalent or similar sentences, and a judgment model to assess the success of a jailbreak attack. We evaluate GPTFuzz against various commercial and open-source LLMs, including ChatGPT, LLaMa-2, and Vicuna, under diverse attack scenarios. Our results indicate that GPTFuzz consistently produces jailbreak templates with a high success rate, surpassing human-crafted templates. Remarkably, GPTFuzz achieves over 90% attack success rates against ChatGPT and Llama-2 models, even with suboptimal initial seed templates. We anticipate that GPTFuzz will be instrumental for researchers and practitioners in examining LLM robustness and will encourage further exploration into enhancing LLM safety.